La factoria d'informació corporativa

Alberto Abelló Gamazo
José Samos Jiménez
Josep Curto Díaz

PID_00189743

Universitat Oberta de Catalunya

Índex

Introducció

Les empreses s'han d'adaptar als canvis del seu entorn (clients, proveïdors, tecnologia, etc.). Per això, cada dia és més important prendre decisions. És en aquest sentit que les dades han guanyat especial importància, fins a esdevenir un bé més de les empreses com podrien ser les matèries primeres, l'energia, el capital o les persones. Cal disposar de totes les dades possibles, tant del negoci com de tot allò que l'envolta, i ser capaç d'analitzar-les de manera eficient per a decidir què és el millor que es pot fer en cada situació.

Generalment, podem considerar que ja tenim la majoria de la informació necessària per a prendre decisions en els sistemes operacionals de l'empresa. En assignatures relacionades amb les bases de dades i l'enginyeria del programari, ja hem vist quines són les característiques d'aquests sistemes operacionals. Dissortadament, com hem vist i continuarem veient més endavant en aquest mòdul, aquest tipus de sistemes no són els més adients per a prendre decisions. Per tant, cal extreure la informació i gestionar-la de manera que es facilitin les tasques d'anàlisi.

Sistemes operacionals

Entenem per sistemes operacionals aquells que ens ajuden en les operacions diàries del negoci, en contraposició amb els sistemes d'anàlisi, que ens ajuden a prendre decisions.

William Inmon va presentar l'any 1998 el que s'anomena *factoria d'informació corporativa*. Es tracta d'un conjunt de components que interactuen per a ajudar a gestionar tots els fluxos de dades des dels sistemes operacionals de l'empresa fins als analistes. El seu objectiu és transformar les dades dels sistemes operacionals (matèries primeres) en informació de suport als analistes (producte elaborat), per a utilitzar-la en els processos de presa de decisions en l'organització. En aquest mòdul veurem els diferents components d'aquesta factoria i com interactuen entre si.

Factoria d'informació corporativa

En anglès, *corporative information factory*. Ho abreujarem com a FIC.

És una factoria o fàbrica d'informació en l'àmbit de tota l'organització o en el corporatiu.

Objectius

Aquest mòdul presenta una arquitectura per a gestionar informació que ajuda a prendre decisions. Es presenten les característiques de cadascun dels elements de l'arquitectura. Amb aquest mòdul assolireu els objectius següents:

1. Veure la necessitat de la factoria d'informació corporativa en la gestió del coneixement per a prendre decisions.

2. Distingir els diferents elements arquitectònics d'una factoria d'informació corporativa.

3. Ser capaços de raonar la necessitat dels diferents sistemes d'emmagatzematge.

4. Saber què són les metadades i quin és el seu paper com a element integrador dels diferents subsistemes.

5. Reconèixer la importància de les metadades en la presa de decisions.

1. Usuaris i fons d'informació dels magatzems de dades

Primerament, en aquest apartat veurem els dos extrems de la cadena d'obtenció d'informació, és a dir, qui són els usuaris de la FIC (què volem) i quines són les fonts d'informació que han de satisfer les seves necessitats (què tenim). Això ens farà pensar sobre què hi ha d'haver al mig per a fer que totes dues coses siguin compatibles.

1.1. Els usuaris

Els sistemes operacionals tenen molts usuaris que accedeixen a molt poques dades, mentre que, pel que fa als sistemes d'anàlisi, els utilitzen molt pocs usuaris que volen veure moltes dades. Recordem que els sistemes operacionals s'utilitzen en el dia a dia de l'empresa. Serveixen per a facilitar tasques rutinàries i repetitives dels oficinistes. Quan parlem de tasques d'anàlisi, les coses esdevenen una mica més complexes i podem identificar **tres tipus diferents d'usuaris**, que podem anomenar *granger*, *explorador* i *turista*. Realment, aquests noms no són gaire importants. El que sí que importa són les característiques de cadascun i els requeriments que tenen.

Lectura recomanada

Podeu trobar els tres tipus d'usuaris extensament explicats en l'obra següent:

W. H. Inmon; C. Imhoff; R. Sousa (1998). *Corporate Information Factory*. EUA: John Wiley & Sons, Inc.

Els analistes tenen requeriments diferents dels que presenten els oficinistes. A més, podem distingir diferents tipus d'analistes amb característiques ben diferents, que la FIC ha de tenir en compte.

1.1.1. Granger

Aquest primer tipus d'usuari porta a terme accessos a la informació absolutament predictibles i repetitius. Regularment troba coses interessants que ajuden al fet que l'empresa funcioni. En tot moment sap què vol i com ho ha d'obtenir, perquè, generalment, repeteix les consultes de manera periòdica. Podríem dir que té la seva parcel·la d'informació i es dedica a conrear-la per a treure'n profit regularment. No accedeix a grans quantitats de dades (ja que mai no surt de la seva parcel·la) i les sol demanar resumides, encara que li pot arribar a interessar veure diferents nivells de detall.

Aquest tipus d'usuari acostuma a utilitzar **eines OLAP**[1]. Aquestes eines estan pensades per a ser utilitzades per personal no informàtic. Són senzilles, entenedores i fan èmfasi en la presentació dels resultats. Mitjançant el model mul-

(1)Sigla de l'expressió anglesa *online analytical processing*, 'processament analític en línia'.

tidimensional (molt proper a la manera d'entendre el negoci d'aquest tipus d'usuaris), aconsegueixen reflectir la complexitat que hi ha en les estructures i relacions de la vida real.

En aquest grup tenim els empleats, els proveïdors i els clients als quals l'organització proporciona serveis informacionals. Actualment, la intel·ligència de negoci operacional, que potencia l'ús d'aquests sistemes en totes les capes de l'organització, permet als usuaris de negoci utilitzar les dades i la informació en els processos de negoci de forma natural, sense haver de sortir de les seves aplicacions. Això es deu al fet que la informació es troba integrada i en qualsevol moment és accessible als processos de negoci, de manera que els usuaris mateixos moltes vegades no són conscients ni del fet que fan servir el magatzem de dades.

Exemple d'anàlisi en línia

Com a exemple de granger, podem pensar en la persona encarregada de fer previsions d'estoc per als magatzems. Aquesta persona segurament voldrà disposar de les dades d'estoc de cada producte durant els darrers anys, i també de les comandes pendents de servir. Basant-se en aquestes dades, haurà de decidir què cal comprar i quan. Si comprés-sim massa o a deshora, l'empresa podria perdre molts diners. No s'ha de confondre aquest analista amb la persona que simplement registra les entrades i sortides del magatzem, el qual no ha de prendre cap decisió.

1.1.2. Explorador

Hi ha altres usuaris analistes que, al contrari que els grangers, tenen uns accessos totalment imprevisibles i irregulars. Passen una gran part del temps sense consultar les dades, planificant o preparant el seu estudi i, quan ho tenen tot a punt, comencen a explorar de cop una gran quantitat de dades, tan detallades com sigui possible. Realment, no saben exactament què busquen fins que ho troben, i els resultats en cap cas no estan garantits. Però, de vegades, troben alguna cosa realment interessant que clarament millora el negoci. Sovint es coneixen com els "usuaris exploradors" (*power users*) de l'organització.

Un usuari explorador acostuma a ser informàtic i/o estadístic, expert en prospecció de dades i per tant amb domini d'eines d'anàlisi estadística. Aquestes eines tracten d'extreure informació oculta (no evident) d'un conjunt de dades. Generalment, són semiautomàtiques (com a mínim demanen alguns paràmetres o que els usuaris validin els resultats) i han d'estar controlades per tècnics especialitzats.

En el context actual, a resultes de la problemàtica existent coneguda com a *big data*, la figura de l'explorador ha evolucionat cap a una nova figura: el **científic de dades** (*data scientific*). Un científic de dades ha de ser capaç d'extreure informació de grans conjunts de dades (en termes del problema de *big data*) d'acord amb un objectiu clar de negoci, no aleatòriament, i posteriorment presentar-la

Big data

Quan parlem de *big data* ens referim al creixement de les dades en volumetria, en velocitat de generació i en variabilitat d'origen i forma.

de manera senzilla a la resta d'usuaris no experts de l'organització. Per tant, es tracta d'un perfil transversal amb coneixements d'informàtica, matemàtiques, estadística, mineria de dades, disseny gràfic, visualització de dades i usabilitat.

Aquest perfil serà clau per a les organitzacions que volen generar avantatges competitius a partir de la informació. En els propers anys, la demanda d'aquest perfil s'incrementarà precisament en aquelles organitzacions que ja tenen en consideració aquest tipus de necessitat i estan desplegant iniciatives d'analítica de negoci, és a dir, en les organitzacions que ja han assolit un nivell de maduresa alt en l'explotació de dades i en la generació d'informació de valor.

Exemples de mineria de dades

Podem utilitzar eines de mineria de dades per a reconèixer patrons de comportament per a detectar fraus (factures, hipoteques o trucades telefòniques impagades), generar regles automàticament per a compondre una cartera de valors invertits en borsa, trobar factors de risc en un postoperatori o descobrir relacions entre les compres de certs productes en el supermercat (per exemple, bolquers i cervesa).

1.1.3. Turista

Hauríem d'entendre aquest darrer tipus d'usuari com un equip format per dues o més persones. D'una banda, tindríem la persona que té una visió global de l'empresa a la qual se li acut la possibilitat de fer un estudi sobre un cert tema. De l'altra, hi hauria un expert en informàtica, coneixedor dels sistemes d'anàlisi de l'empresa, encarregat d'esbrinar si l'estudi és factible amb les dades i eines disponibles o no.

Aquest equip mirarà dades sense seguir cap patró d'accés i rarament mirarà dos cops les mateixes dades. Per tant, tampoc no en podem conèixer els requeriments *a priori*. A més de les dades, també estarà especialment interessat a consultar les metadades. Les eines que utilitzaran els turistes són **navegadors o cercadors** (tant de dades com de metadades) i el resultat de la seva feina seran projectes que duran a terme els grangers o els exploradors.

Les metadades

Les metadades són dades sobre les dades. Les podeu veure més detalladament en l'apartat "Les metadades" d'aquest mateix mòdul.

Un usuari turista és, en definitiva, un usuari casual de la informació.

1.2. Les fonts d'informació i les seves dades

La primera pregunta que ens hem de fer és si ja tenim la solució als problemes que presenten aquests usuaris i eines d'anàlisi. Actualment, el que tenen les empreses és un conjunt d'aplicacions independents, posades en marxa en diferents moments, que donen resposta a diferents requeriments. La gran majoria d'aquestes aplicacions només estan concebudes per a donar suport al procés de negoci (per exemple, la gestió de personal, comptabilitat, etc.). Cap d'aquestes aplicacions no es va dissenyar per a ser utilitzada pels analistes.

L'espaiament dels moments de desenvolupament de cada aplicació, les diferències de pressupostos i requeriments i la manca de planificació fan que trobem més heterogeneïtats de les que volíem entre els sistemes operacionals. Si els analistes volguessin accedir directament a les dades d'aquestes fonts d'informació, el primer que haurien de fer seria **superar aquestes heterogeneïtats entre les aplicacions**.

Tipus d'heterogeneïtats

Podem trobar tant heterogeneïtats semàntiques (el mateix tipus d'informació representat de maneres diferents), com de sistemes (per exemple, maquinari diferent, sistema operatiu diferent o simplement sistema de gestió de base de dades –SGBD– diferent).

Encara que els analistes fossin capaços d'accedir als múltiples sistemes operacionals alhora, cal tenir present que cada un d'aquests ha estat dissenyat per a resoldre un cert problema de manera eficient. Una de les implicacions d'això és que no guarden més dades de les necessàries per a resoldre el problema corresponent, ja que això empitjoraria la seva eficiència. Concretament, **no guarden dades històriques**, si no cal. Aquest tipus de dades són imprescindibles per a tenir una referència a l'hora de prendre decisions. Una implicació directa d'aquesta necessitat de dades històriques és el gran volum de dades que han de gestionar els sistemes d'anàlisi.

Exemple de dades necessàries per a una anàlisi de telefonia a Catalunya

Imaginem que utilitzem 4 bytes per a codificar l'origen d'una trucada i 4 més per a codificar la destinació. A més, també podem codificar el moment en què es fa la trucada amb 4 bytes i la seva durada amb 2. En total, per a cada trucada només necessitarem guardar 14 bytes. Pensem que la mitjana de trucades per persona i dia és aproximadament de deu i que a Catalunya hi viuen sis milions de persones. Si volguéssim fer un estudi dels darrers tres anys, tindríem que ens cal el següent:

3 anys × (365 dies/any) × (6×10^6 persones) × (10 trucades/persona i dia) × (14 bytes/trucada) = 10^{12} bytes = 10^9 kB = 10^6 MB = 10^3 GB = 1 TB.

Això vol dir que necessitaríem deu discos de 100 GB per a guardar tota aquesta informació i aproximadament estaríem dues hores i mitja llegint totes les dades (si assumim una velocitat de lectura de 100 MB/segon). Aquest temps de resposta resulta clarament inadmissible per a qualsevol persona que faci una consulta interactiva. Penseu que les tècniques d'indexació habituals tampoc no serveixen de gaire quan el que volem consultar és la suma, la mitjana, el mínim o el màxim de la durada de les trucades.

En contraposició als sistemes OLAP, podríem identificar els sistemes operacionals amb els processaments transaccionals en línia (OLTP, *on-line transactional processing*). Aquests sistemes transaccionals estan pensats per a obtenir dades. Per tant, els és essencial evitar la introducció de dades errònies i permetre accessos concurrents de manera aïllada, sense interferències. Així, doncs, estan dissenyats per a manipular una gran quantitat de petites transaccions que impliquen modificacions de les dades.

En canvi, per als analistes, la disponibilitat de les dades és molt més important que l'aïllament. Les seves consultes són molt més complexes i hi involucren moltes dades. Aquests analistes no poden esperar un cert conjunt de dades bloquejat per algú que les modifiqui, perquè, amb la gran quantitat de dades que consulten, la probabilitat que algú n'estigués modificant alguna seria massa alta. A més, **els analistes només volen fer consultes** (estem en un

entorn només de lectura, *read only*) de manera que totes les precaucions del món transaccional (pensades per a entorns de lectura/escriptura, *read/write*) són totalment innecessàries.

Els sistemes operacionals demanen un bon rendiment en l'execució de transaccions que sempre han de deixar la base de dades en un estat consistent, mentre que els sistemes d'anàlisi requereixen executar consultes complexes que retornin dades precises en un temps de resposta baix. Intentar de compatibilitzar requeriments només de lectura o de lectura/escriptura seria dolent per a tots dos entorns. Això vol dir que no podem utilitzar els sistemes operacionals, sinó que hem de crear sistemes independents per als analistes.

Tot i que les dades dels sistemes operacionals que té l'empresa ens siguin molt interessants, aquests sistemes no compleixen els requeriments dels analistes. Cal definir un sistema que aprofiti aquestes dades i que satisfaci les necessitats d'aquests usuaris de manera adequada.

1.3. El magatzem de dades

La solució a les necessitats dels analistes és construir una base de dades d'anàlisi, que anomenarem **magatzem de dades**, partint de les bases de dades operacionals, però que funcioni independentment d'aquestes.

Vegeu també

Vegeu la definició de *magatzem de dades* que apareix en el mòdul "Introducció a l'emmagatzematge de dades".

2. Els magatzems de dades departamentals

Construir un magatzem de dades és molt costós, a més de tenir uns requeriments de rendiment difícils d'aconseguir. La solució per a obtenir un temps de resposta baix és disposar de diferents magatzems només amb informació parcial del negoci (només la part que interessi a un cert departament o conjunt de persones).

Aquests **magatzems de dades departamentals**[2] normalment estaran dissenyats seguint el model multidimensional, cosa que facilita la millora en el rendiment, mitjançant tècniques específiques d'emmagatzematge de les dades. A més, per a no sobrecarregar els sistemes amb dades innecessàries, només contenen dades històriques dins el període de temps que sigui estrictament necessari.

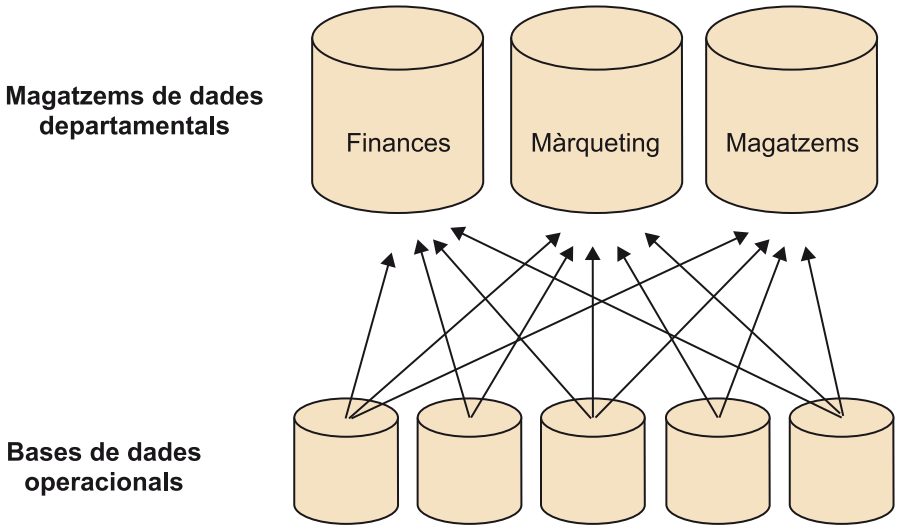
[2]En anglès, *data marts*.

Vegeu també

Veurem el model multidimensional en el mòdul "Disseny multidimensional" d'aquesta assignatura.

Exemple de tècnica d'emmagatzematge per a millorar el temps de resposta

Com a exemple de tècnica per a millorar el temps de resposta, podem pensar en la preagregació. Aquesta tècnica consisteix a guardar els resultats de les funcions d'agregació (suma, mitjana, mínim, etc.) ja calculats per quan l'usuari els demani. Això vol dir que hem de conèixer (o imaginar) quines consultes voldrà fer, per a calcular prèviament els resultats, de manera que el càlcul no s'hagi de fer en el moment concret en què se sol·liciten.



Com podeu veure en la figura superior, per a cada grup d'usuaris o departament que ho requereixi construïm un d'aquests magatzems, que només integra les dades de les fonts d'informació que calguin per a satisfer les necessitats concretes del seu grup d'usuaris (cosa que també en facilita el funcionament). Aquestes dades es modelen seguint la visió de la realitat que tingui el departament corresponent i no cal que es consensuï amb tota l'empresa.

Un altre avantatge dels magatzems de dades departamentals és que no els cal tenir les dades amb el màxim nivell de detall. Per exemple, si els analistes només volen veure les dades mensuals, no és necessari emmagatzemar les dades diàries. Així, no caldria emmagatzemar les vendes diàries de l'empresa, sinó solament el total que s'ha venut durant un mes, fet que representa un estalvi d'espai clar.

Tenir molts magatzems de dades petits permet d'abaratir costos, ja que són més econòmics que un de gran que satisfaci les necessitats de tothom alhora. A més, fent-ho així, facilitem la configurabilitat. Finalment, també és més fàcil controlar tant els costos (que s'imputaran al departament corresponent), com els accessos, processos i configuració del sistema (que correspondran a un conjunt d'usuaris molt restringit).

Problemes d'afinació

En qualsevol base de dades, com més usuaris tenim més difícil es fa compatibilitzar tots els requeriments per a aconseguir el rendiment òptim.

Els magatzems de dades departamentals guarden una història parcial de les dades que interessen a un cert departament. Estan dissenyats per a obtenir un temps de resposta bo davant les consultes d'un cert conjunt d'analistes.

3. El magatzem de dades corporatiu

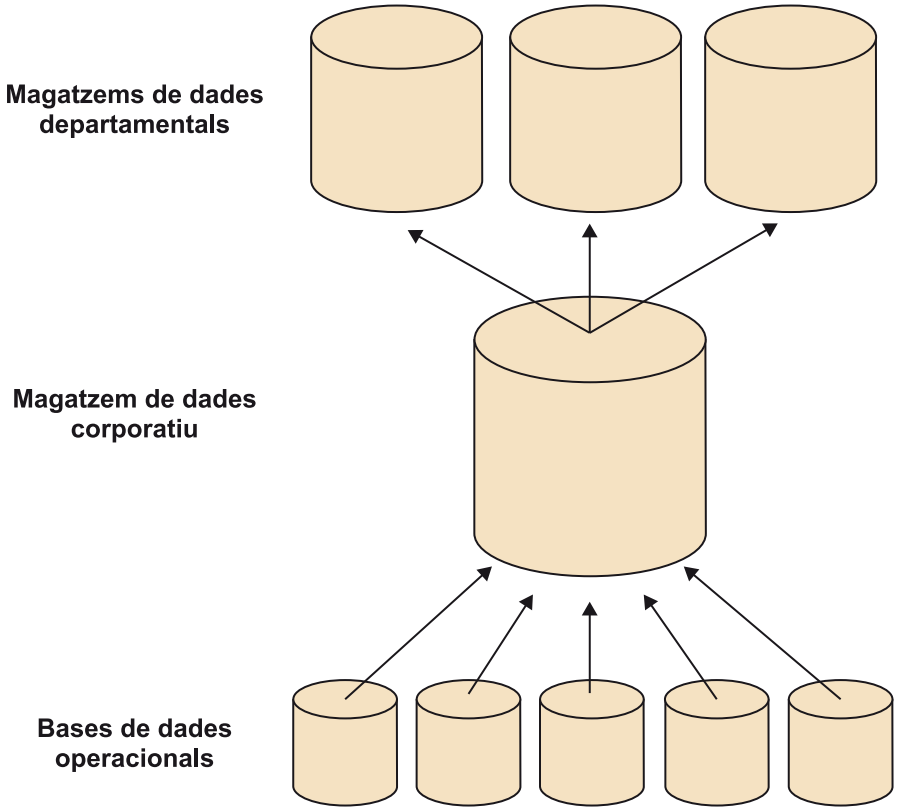
Magatzem de dades corporatiu

Acostumem a referir-nos al magatzem de dades corporatiu amb el terme anglès *data warehouse*.

Tenir múltiples magatzems de dades departamentals independents genera problemes a llarg termini, tot i que són més econòmics i fàcils de construir a curt termini. El primer problema és que, com podeu veure en la figura anterior, tenim processos independents d'integració i transformació per a cada magatzem de dades departamental. A més, on guardem la informació que actualment no interessa a cap departament? No tenim cap lloc on la puguem guardar i no la podem llençar. Cal tenir un magatzem de dades corporatiu que guardi tota la història de totes les dades i sempre amb el màxim nivell de detall possible. No obstant això, els magatzems de dades departamentals encara són necessaris.

El magatzem de dades corporatiu no és apropiat per als usuaris finals, perquè està dissenyat per a gestionar i integrar grans quantitats de dades que, juntament amb l'excés d'usuaris, degraden el temps de resposta. No es pot dissenyar per a afavorir un grup d'usuaris concret, sinó que ha de servir a tots alhora de la millor manera possible.

Així, com es pot veure en la figura següent, el magatzem de dades corporatiu resulta d'un procés d'integració i transformació de totes les fonts de dades únic i complex, que estudiarem detalladament en l'apartat corresponent d'aquest mòdul. Els magatzems de dades departamentals ara s'obtenen simplement com a resultat d'un procés de transformació a partir del magatzem corporatiu.



El magatzem de dades corporatiu guarda tota la història de totes les dades de l'empresa integrades. Està dissenyat per a emmagatzemar-les eficientment.

La taula següent resumeix les diferents característiques dels dos tipus de magatzem de dades que hem vist fins ara:

Característica	Magatzem de dades	
	Departamental	Corporatiu
Temàtica	Específica	Genèrica
Fonts de dades	Poques	Moltes
Grandària	Gigabytes	Terabytes
Temps de desenvolupament	Mesos	Anys
Model de dades	Multidimensional	Relacional

Primerament, el magatzem de dades corporatiu ha de ser genèric i ha de guardar dades de tota l'empresa seguint una visió consensuada del negoci. Per contra, els magatzems de dades departamentals són absolutament específics. Només contenen les dades que demana un cert conjunt d'usuaris, les guarden segons la concepció que aquests tenen del negoci i estan optimitzats per a obtenir un bon rendiment davant les tasques d'anàlisi que aquests volen fer.

Realment, els magatzems de dades departamentals no s'alimenten directament de les fonts de dades, sinó del magatzem de dades corporatiu. No obstant això, per transitivitat i sense oblidar aquesta puntualització, també podem estudiar la diferència que hi ha entre les fonts de dades dels magatzems departamentals i dels corporatius. Com ja hem vist, els magatzems de dades departamentals només contenen les dades que interessen a un cert grup d'analistes. Per tant, només guardaran dades que provinguin de les fonts de dades corresponents. En canvi, el magatzem de dades corporatiu, com que conté totes les dades que interessen o poden arribar a interessar als analistes de qualsevol departament, s'ha d'alimentar de la unió de totes les fonts d'informació de tots els magatzems departamentals, a més d'aquelles fonts de dades que avui no interessen a cap analista, però que potencialment poden arribar a interessar a algú.

Seguint el mateix raonament que acabem de fer per a les fonts d'informació, també podem veure que els volums de dades que contenen tots dos tipus de magatzems de dades han de ser d'ordres de magnitud diferents. Generalment, podem considerar que un magatzem de dades departamental amb dades només d'un cert conjunt de temes i el qual no les conté amb el màxim nivell de detall pot ocupar uns quants (possiblement molts) gigabytes (10^9 bytes) de dades. Per contra, el magatzem de dades corporatiu, que alhora ha de contenir les dades de tots els magatzems departamentals i no pot perdre detall (ha de guardar la informació tan detallada com sigui possible, per si algun dia algú la necessita), ocuparà un nombre de terabytes (10^{12} bytes) de dades.

Fent una simple regla de tres, podem deduir que, si un tipus de magatzem conté moltes més dades que l'altre, ha d'integrar més fonts de dades i ha de ser molt més genèric, llavors trigarem molt més temps a desenvolupar-lo i hi haurem d'invertir molts més recursos (tant econòmics com humans). Generalment, podem considerar que un projecte per a construir un magatzem de dades departamental dura uns quants mesos, mentre que desenvolupar el magatzem de dades corporatiu de l'empresa és un procés que dura anys.

El magatzem de dades corporatiu acostuma a estar implementat sobre un SGBD relacional simplement per les prestacions que ofereixen aquests sistemes en la gestió de grans volums de dades, no perquè aquests sistemes estiguin especialment concebuts per a això. I encara més, el rendiment dels sistemes transaccionals acostuma a ser especialment dolent per a tasques d'anàlisi, si no s'estenen amb mecanismes específics d'anàlisi (per exemple, nous tipus d'índexs, càrrega massiva de dades, etc.). En canvi, els magatzems de dades

Vegeu també

Veurem els diferents sistemes multidimensionals (ROLAP, MOLAP, HOLAP, etc.) en el mòdul "Disseny multidimensional".

departamentals, sense un requeriment tan gran respecte al volum de dades i valorant molt més el temps de resposta, s'acostumen a implementar sobre sistemes que es basen en el model multidimensional.

Els sistemes operacionals estan dissenyats per a respondre bé davant de petites transaccions que modifiquen les dades. El model relacional ofereix tota la teoria de la normalització per a aconseguir que els SGBD tinguin un bon rendiment amb aquest tipus d'accessos, de manera que una modificació afecti una única taula i s'evitin les anomalies d'inserció, actualització i esborrament. Però en el cas del magatzem de dades corporatiu, ens hem de plantejar si, encara que utilitzem un SGBD relacional, realment cal normalitzar el nostre esquema. La càrrega de dades es produeix de manera massiva, tota de cop i en el moment que els usuaris no fan consultes. Què té això en comú amb les petites transaccions d'alta, baixa, modificació i consulta de les bases de dades operacionals?

4. El magatzem de dades operacional

Malauradament, és possible que, amb els magatzems de dades departamentals i el corporatiu encara no tinguem cobertes totes les necessitats d'informació de l'empresa. A causa del seu volum de dades i de les tècniques d'implementació que s'utilitzen, el magatzem de dades corporatiu (i conseqüentment els departamentals que s'actualitzen a partir d'aquest) no es pot tenir constantment actualitzat (només s'acostuma a actualitzar durant les nits o els caps de setmana). D'altra banda, els seus usuaris tampoc no ho requereixen, ja que estan més interessats en les dades històriques que en les actuals. Però hi pot haver altres usuaris que també demanin dades integrades i que les vulguin completament actualitzades. Encara necessitem un altre tipus de repositori d'informació.

El magatzem de dades operacional és una estructura a cavall entre el món operacional i el de la presa de decisions. Està orientat al tema i integrat com qualsevol magatzem de dades, però en aquest cas no conté cap tipus d'informació temporal.

Lectura complementària

Podeu trobar molta més informació sobre el magatzem de dades operacional en el llibre següent:

W. Inmon; C. Imhoff; G. Batas (1996). *Building the Data Warehouse* (2a. ed.). EUA: John Wiley & Sons, Inc.

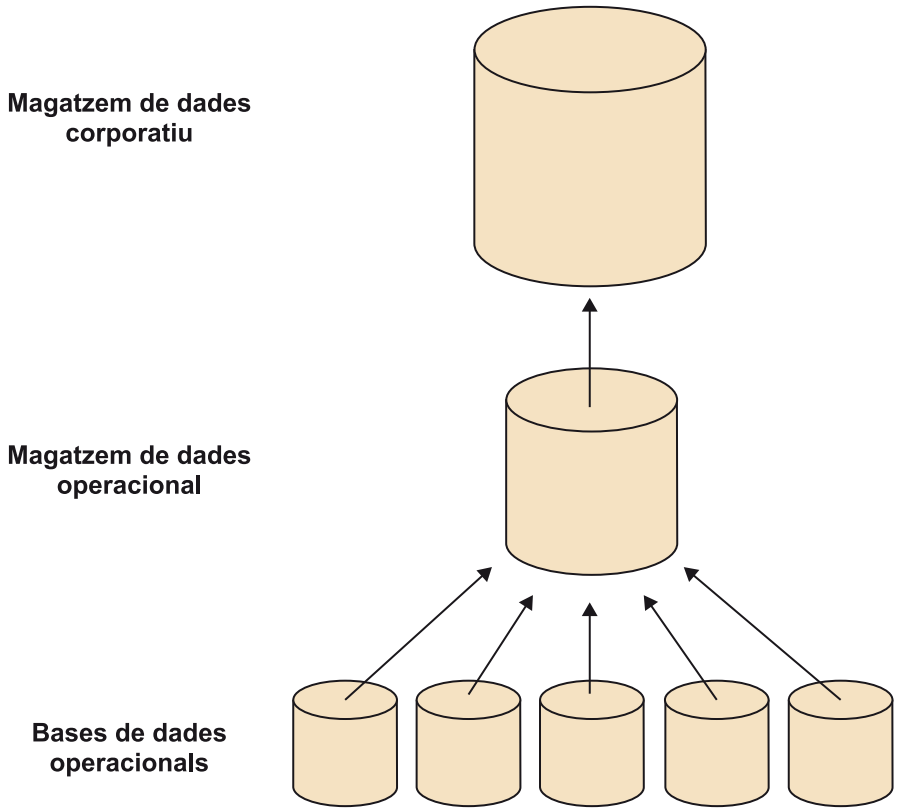
L'aparició d'aquest repositori és donada per la típica ponderació entre volum de dades i velocitat del sistema. Fins ara, en els altres magatzems, el que volíem era tenir absolutament qualsevol dada que poguéssim arribar a necessitar per a prendre una decisió. Com a conseqüència d'aquest requeriment, el temps de resposta pot arribar a degradar-la i, en qualsevol cas, ens veiem obligats a renunciar a tenir les dades constantment actualitzades. En aquest cas, valorem més el fet que les dades sempre estiguin actualitzades, que no que les tinguem totes. Per tant, renunciem a tenir dades històriques i tenim un repositori volàtil.

Aquest és el preu que s'ha de pagar per a reduir el volum de dades i poder-lo mantenir constantment actualitzat. D'aquesta manera, el magatzem de dades operacional i el corporatiu es complementen: el corporatiu guarda totes les dades històriques, però no està actualitzat sempre, i l'operacional sempre està actualitzat, però no conté dades històriques.

A més de permetre l'accés a dades operacionals integrades actualitzades, com podem veure en la figura següent, també facilita la construcció del magatzem de dades corporatiu. Es pot veure com una estratègia de dividir i vèncer. En comptes d'aconseguir les quatre característiques del magatzem de dades corporatiu de cop, primer n'aconseguim dues mitjançant el magatzem de dades

(3)En anglès, *back-up*.

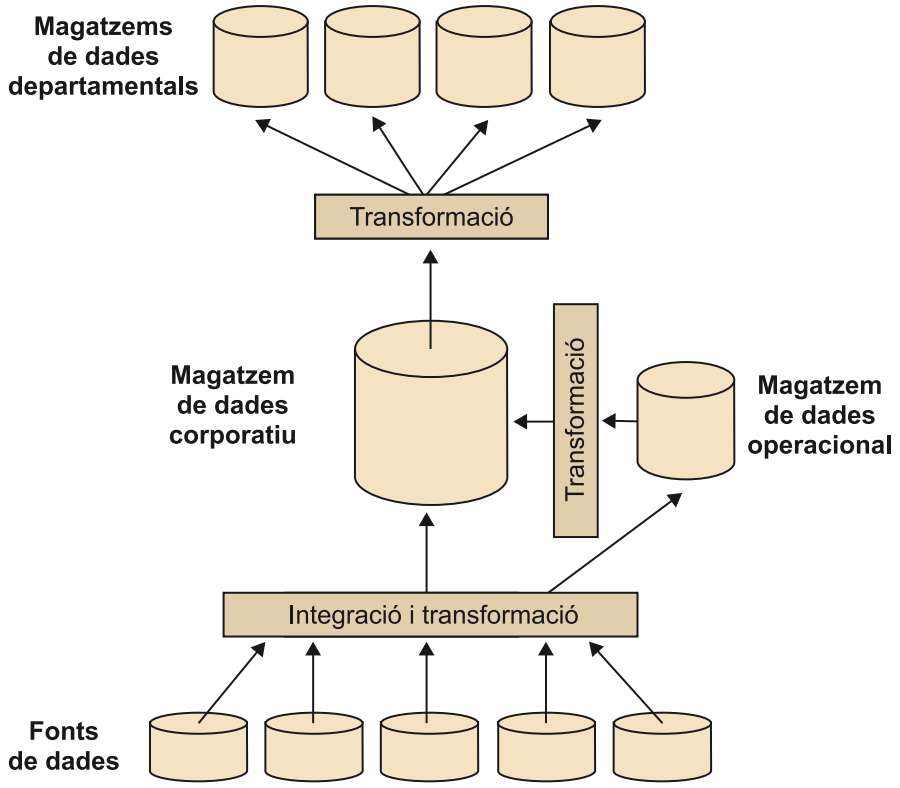
operacional (orientació al tema i integració) i en un segon pas afegim la temporalitat (historicitat i no-volatilitat). El pas del magatzem de dades operacional al corporatiu pot ser tan senzill com fer un bolcatge[3] de les dades.



5. El component d'integració i transformació

Com hem vist en l'apartat "Usuaris i fonts d'informació dels magatzems de dades", els sistemes operacionals dels quals disposen les organitzacions generalment no compleixen els requeriments dels analistes. Com a solució, s'ha definit el concepte de *magatzem de dades*, tant en l'àmbit departamental com en el corporatiu, segons les característiques de les seves dades, que el diferencien dels sistemes operacionals.

Tanmateix, les dades dels magatzems de dades s'obtenen a partir dels sistemes operacionals de l'empresa, i també a partir de fonts externes. Per les seves característiques diferents pel que fa a estructura i organització, les dades obtingudes de les fonts no es poden utilitzar directament en el magatzem de dades, sinó que s'han d'adaptar als seus requeriments en aquests aspectes.



Flux de dades mitjançant el component d'integració i transformació.

La missió del **component d'integració i transformació** consisteix a obtenir les dades per als diferents magatzems de dades de l'organització.

Els sistemes operacionals

Els sistemes operacionals són la porta principal d'entrada de dades a l'organització, encara que aquesta també pot utilitzar dades externes obtingudes de diferents fonts (informes especialitzats, articles de premsa, etc.). En els últims temps la font principal de dades externes per a les organitzacions sol ser Internet.

Els termes *integració* i *transformació*

El component d'integració i transformació se sol referenciar mitjançant les sigles en anglès: I&T (*integration and transformation*). També és freqüent trobar-lo referenciat com a *procés ETL* (*extraction, transformation and loading*). Perquè el seu nom fos totalment descriptiu s'hauria d'anomenar *component d'obtenció, transformació, depuració, integració, transport i càrrega*. Aquí farem servir la terminologia definida a Inmon, Imhoff i Sousa (1998).

Integració i transferència de les dades

Com podem veure en la figura, les dades obtingudes de les fonts de dades es carreguen en el magatzem de dades operacional i des d'allà es transfereixen al magatzem de dades corporatiu, encara que algunes també s'hi poden carregar directament. Les dades ja integrades al magatzem de dades corporatiu es transformen i carreguen als diferents magatzems de dades departamentals.

Originalment les dades s'obtenen a partir dels sistemes operacionals i altres fonts de dades, i s'han de transformar, depurar i integrar segons l'estructura dels esquemes dels magatzems de dades també s'ha de transportar i carregar perquè es puguin utilitzar en els diferents magatzems de dades de l'organització.

A diferència dels magatzems de dades, l'element principal dels quals és la base de dades, l'element principal del component d'integració i transformació és el programari encarregat de dur a terme la missió descrita.

Tant les fonts de dades com els diferents magatzems de dades es poden trobar en plataformes diferents, per tant, el component d'integració i transformació tindrà elements en les diferents plataformes en què hi hagi la resta de components de la FIC.

El component d'integració i transformació està format per programari que s'executa en les diferents plataformes en què funcionen la resta de components de la FIC.

En aquest apartat estudiarem les activitats esmentades del component d'integració i transformació.

5.1. Obtenció de les dades

A partir de les fonts de dades hem d'obtenir les dades requerides pels magatzems de dades de la FIC. Els magatzems de dades guarden la història de les dades per a permetre d'analitzar la seva evolució; els sistemes operacionals generalment tan sols en guarden una imatge. És a dir, partint de sistemes que mantenen una "imatge" de les dades, hem d'obtenir aquelles que són necessàries per a formar la seva història, que podríem representar com una "pel·lícula" de les dades, entenent-la com una seqüència d'imatges. Hem d'obtenir les actualitzacions que es produeixen sobre les dades per a anar muntant aquesta pel·lícula[4].

Pel·lícula

Encara que freqüentment es parli de *pel·lícula* per a donar una visió de les dades en un magatzem de dades, aquestes no estaran formades necessàriament per una successió d'imatges, sinó que, entre d'altres possibilitats, poden consistir en una imatge inicial i una sèrie de diferències de les successives imatges obtingudes.

En l'obtenció de les dades d'un magatzem de dades es distingeixen dues fases:

- Obtenció de les dades de la imatge inicial.
- Obtenció de les dades per a les actualitzacions.

(4) *Fotografia* o *imatge*; en anglès, s'anomena *snapshot*.

Les fonts de dades generalment mantenen les seves dades estructurades segons la utilització que se'n faci. A causa de la falta d'integració entre les diferents aplicacions, és freqüent trobar dades replicades entre si. És a dir, per a cada dada considerada podem trobar diferents fonts disponibles.

El primer pas en l'obtenció de les dades consisteix a determinar, d'entre totes les fonts possibles, quina és la més adequada per a cada una de les dades requerides: s'ha de determinar el que Inmon anomena el **sistema de registre**.

Obtindrem cada dada de la font o fonts que més bé s'adaptin als requeriments de qualitat, precisió, estructura o disponibilitat dels magatzems de dades. Així mateix, per a cada dada haurem de determinar quina font és la més adequada per a obtenir la imatge inicial i d'on obtindrem les successives actualitzacions.

Actualment, el patró de creixement de les dades en el context empresarial ha canviat per motius diversos. D'una banda, per l'aparició de múltiples dispositius nous que generen dades de valor noves per a les organitzacions (per exemple, sensors distribuïts en una ciutat per a monitorar l'eficiència del trànsit o els sistemes de distribució d'aigua, gas o electricitat). De l'altra, l'usuari cada cop presenta un comportament més actiu a través de les xarxes socials, el comerç electrònic i els nous dispositius intel·ligents com telèfons intel·ligents (*smartphones*) i tauletes (*tablets*), i per tant podem dir que esdevé el generador de dades principal. Com a conseqüència de tot això, la mida del fitxer es veu reduïda comparativament i la quantitat de dades en trànsit genera una ombra digital de valor alt per a les organitzacions (per exemple, en els sistemes de recomanació).

En aquest nou context, la informació de valor per a una organització no sempre es troba en els sistemes transaccionals i, sovint, les dades no són estructurades. Segons el grau d'estructuració (i per extensió, la dificultat d'extracció de la informació) de les dades, les podem classificar en els tipus següents:

1) Dades estructurades: es caracteritzen per tenir una estructura coneguda i s'emmagatzema principalment en bases de dades relacionals. La manipulació de les dades es fa per mitjà de gestors de bases de dades, i les consultes per mitjà d'SQL.

2) Dades semiestructurades: es troben encapsulades en fitxers semiestructurats com XML[5] o SGML[6]. En aquesta situació és possible treballar amb el context de negoci, cosa que proporciona gran valor a les organitzacions. Actualment existeixen bases de dades especialitzades en XML per a manipular aquest tipus de dades i també tècniques como *web-mining* (mineria de dades aplicada al web) que permeten recuperar informació de pàgines web.

(5)De l'anglès, *extensible markup language*.

(6)De l'anglès, *standard generalized markup Language*.

3) Dades no estructurades: encapsulades en objectes sense una estructura predefinida (àudio, vídeo, PDF o Word) que requereix l'ús de tècniques especials com *text-mining* (mineria de dades aplicada a fitxers de text) o *information retrieval* (tècniques, sovint estadístiques, aplicades a trobar informació relacionada amb un concepte en fitxers).

5.1.1. Obtenció de les dades de la imatge inicial

Generalment, els sistemes operacionals només guarden una imatge de les seves dades o bé una història reduïda d'aquestes. Aquesta imatge és la que s'ha d'obtenir per a traspassar-la als magatzems de dades.

Si les diferents imatges emmagatzemades en els sistemes operacionals s'han anat perdent a mesura que s'han fet modificacions, només podrem disposar de la història que emmagatzemem a partir del moment en què es construeixin els magatzems de dades.

En alguns casos, per diferents motius (per exemple, per motius legals) hi pot haver una història més extensa de les dades, de vegades fora dels sistemes operacionals, encara que obtinguda a partir d'aquests. En cada cas haurem de valorar si és útil per als analistes disposar en els magatzems de dades de la història que hi havia abans; en cas positiu, en lloc de partir de la imatge inicial partiríem d'una pel·lícula inicial.

Dades històriques en un banc

En un banc, el sistema operacional de gestió de moviments dels comptes només guarda les dades dels últims dotze mesos. Les dades dels mesos anteriors fins a un total de cinc anys s'han d'emmagatzemar per motius legals. Tanmateix, aquests moviments històrics no s'emmagatzemen en el sistema operacional, sinó que mensualment s'extreuen de la base de dades del sistema i s'emmagatzemen en un mitjà d'emmagatzematge més econòmic. Encara que aquestes dades romanen accessibles dins l'organització, només s'hi accedeix de manera puntual, per motius operacionals, no per a analitzar-les.

Per a obtenir la imatge inicial haurem de desenvolupar un conjunt d'aplicacions d'obtenció de les dades de les fonts de dades, les quals generalment s'executaran una sola vegada.

5.1.2. Obtenció de les dades per a les actualitzacions

Una vegada tenim la imatge inicial de les dades en els magatzems de dades, repetitivament, segons les necessitats dels analistes, haurem d'obtenir les modificacions fetes en les fonts de dades. D'aquesta manera anirem formant la pel·lícula de les dades que oferirem als usuaris perquè l'analitzin.

Per a obtenir les actualitzacions haurem de desenvolupar un conjunt d'aplicacions d'obtenció de les dades de les fonts de dades, les quals s'executaran freqüentment.

5.2. Transformació, depuració i integració de les dades

Quan ja hem obtingut les dades de les diferents fonts:

- Cada conjunt de dades pot tenir una estructura diferent depenent de la font de la qual procedeixi. Les hem de transformar per adaptar-les a l'estructura de l'esquema del magatzem de dades en què s'emmagatzemaran.

- Hem de depurar els errors o conflictes que puguem trobar dins les dades de cada una de les fonts.

- Hem d'integrar les dades depurant errors o conflictes entre dades de fonts diferents.

Com a resultat d'aquest procés, obtindrem un conjunt de dades directament utilitzable per a actualitzar el magatzem de dades corresponent.

A continuació, comentarem alguns detalls d'aquestes operacions per separat. Això no significa que es facin de manera seqüencial, sinó que es poden combinar o intercalar algunes d'aquestes segons les necessitats.

5.2.1. Transformació de les dades

Les transformacions que cal fer sobre les dades poden ser molt variades. Entre les més freqüents trobem les següents:

- Canviar el format o el tipus de les dades (per exemple, els camps de data).

- Canviar la codificació (per exemple, EBCDIC a ASCII).

- Reestructurar els camps (per exemple, fusionar o dividir camps, canviar el seu ordre relatiu).

- Canviar les unitats o codis de representació (per exemple, canvis de moneda).

- Canvis en el grau d'agregació (per exemple, calcular les vendes mensuals a partir de les diàries).

- Calcular camps derivats (per exemple, calcular l'edat a partir de la data de naixement).

- Afegir informació temporal (per exemple, període de validesa de les dades).

Una de les transformacions que generalment sempre s'ha de fer és l'última assenyalada, la d'afegir a les dades informació temporal: s'haurà d'afegir la informació sobre el període de validesa de les dades o el moment en què s'hagi registrat la modificació (o en què s'hagi detectat), segons sigui requerit pel magatzem de dades corresponent. D'aquesta manera seqüenciem les imatges obtingudes per a anar formant la pel·lícula que emmagatzema el magatzem de dades.

5.2.2. Depuració de les dades

L'objectiu de depurar les dades obtingudes de les diferents fonts és millorar-ne la qualitat. Algunes de les incidències més comunes que es produeixen són les següents:

- Detectar i corregir valors inconsistents (per exemple, un atribut edat amb un valor de tres-cents cinquanta).

- Afegir valors per defecte als camps amb valors no definits. Generalment, es fa d'acord amb criteris marcats pel magatzem de dades al qual es destinen les dades, segons la font de dades. El valor subministrat pot ser constant, calculat o, en alguns casos, pot interessar deixar-lo sense definir.

- Detectar i corregir informació duplicada. De vegades és difícil de detectar, ja que es tenen diferents representacions del mateix valor (per exemple, diferents maneres d'escriure el nom d'un carrer en les dades del domicili). Serà més freqüent trobar informació duplicada entre diferents fonts de dades, però també la podem trobar dins una mateixa font.

5.2.3. Integració de les dades

Ateses les dades obtingudes de diferents fonts, les hem d'integrar entre si, i també amb les dades del magatzem de dades al qual es destinen.

> El procés d'integració serà diferent depenent de si fem la càrrega inicial del magatzem de dades, o bé una actualització d'aquest.

A més del volum de dades que cal tractar, la diferència principal rau en el fet que en les actualitzacions, per a fer la integració, podem fer servir les correspondències entre les dades de les fonts i les del magatzem de dades prèviament

establertes en la càrrega inicial o en actualitzacions anteriors. Generalment, en el procés de càrrega inicial es farà una integració de totes les dades prèvia a fer la càrrega en el magatzem de dades. D'altra banda, quan es fa l'actualització, és possible que no estiguin disponibles les dades de totes les fonts alhora i interessi integrar les dades de les diferents fonts per separat en el magatzem de dades.

El problema principal amb què ens trobem consisteix a detectar quines dades representen el mateix concepte.

Si les diferents fonts de dades utilitzen com a clau el mateix camp de l'entitat (per exemple, NIF), es poden relacionar sense dificultat, excepte per errors en les dades. El problema sorgeix quan cada font de dades empra la seva clau (per exemple, un codi generat) i no hi ha camps comuns que puguin servir com a clau alternativa per a establir relacions entre si o, si n'hi ha, els seus valors es representen de manera diferent entre les fonts.

Durant el procés d'integració es transformaran les dades per a homogeneïtzar la seva representació i s'eliminarà la informació duplicada.

S'hauran d'establir els procediments adequats per a propagar les correccions fetes fins als sistemes operacionals dels quals procedeixen les dades. Aquestes seran especialment rellevants després d'obtenir les dades per a la càrrega inicial del magatzem de dades, però també s'hauran de tenir en compte les fetes en cadascuna de les actualitzacions de les dades.

Dades depurades

Si hem dedicat un esforç considerable per a integrar les dades de clients de diferents sistemes operacionals, depurant-les i eliminant duplicats, el que és raonable és utilitzar les dades depurades en els sistemes operacionals, en lloc de continuar utilitzant-les amb errors. Per això, s'haurà de definir un sistema per a propagar les correccions fetes en les dades des del component d'integració i transformació fins als sistemes operacionals dels quals procedeixen.

5.3. Transport i càrrega de les dades

Quan ja hem obtingut les dades, s'han de transportar des de les diferents plataformes de les fonts fins a les plataformes dels magatzems de dades als quals s'incorporaran. També és necessari transportar-les entre diferents magatzems de dades (vegeu la figura anterior).

El component d'integració i transformació també s'encarrega de transportar les dades entre les diferents plataformes i carregar-les en les bases de dades corresponents.

Tant en el transport com en la càrrega de les dades s'ha de distingir entre el procés de càrrega inicial, que s'executarà una sola vegada, i el procés d'actualització, executat freqüentment. El transport i la càrrega inicials es poden resoldre de manera puntual, sense que hi hagi la necessitat de dedicar recursos permanents amb aquesta finalitat. Només s'hauran de tenir disponibles permanentment els recursos per a fer les actualitzacions de les dades.

6. Les metadades

Les metadades no són un element específic de la FIC, apareixen en molts contextos del món del programari. La definició més freqüent que hi ha del concepte de metadada està basada en la seva etimologia[7]: "Les metadades són dades sobre dades". Les dades generalment representen característiques de les entitats que modelen; en el cas de les metadades, representen característiques d'altres dades que en faciliten l'administració i ús. És a dir, el que diferencia una dada d'una metadada més que la seva estructura o contingut és el seu propòsit i ús.

(7) *Meta* en grec significa 'sobre'.

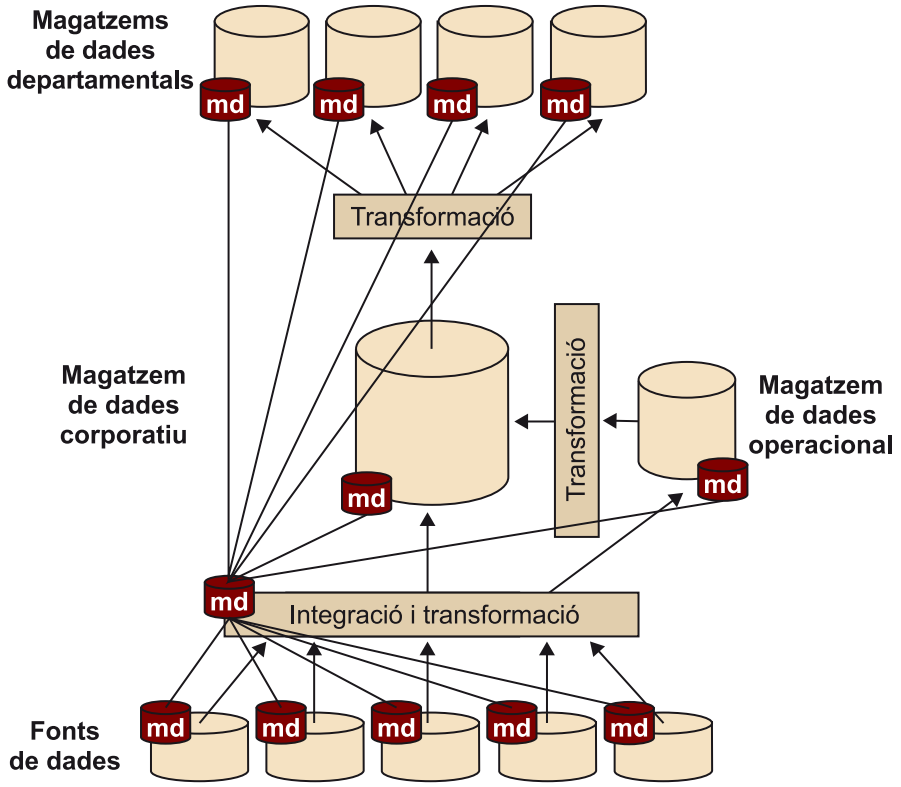
Atès un conjunt de dades, les metadades sobre aquestes descriuen les seves característiques (per exemple, format, origen, ús, etc.). Aquestes metadades són dades i al seu torn podem tenir altres metadades que descriguin les seves característiques (metadades sobre metadades), i així successivament.

En aquest apartat comencem revisant l'ús de les metadades en la FIC. A continuació es presenten diferents tipus de metadades segons l'ús que se'n fa. També s'analitza la manera en què es creen les metadades, i també els estàndards definits per a permetre de compartir-ne entre diferents components. La secció acaba comentant la necessitat de fer servir diferents versions de metadades en la FIC.

6.1. Metadades i els components de la FIC

En la FIC es produeix un flux de dades des de les fonts d'aquestes fins als analistes. Aquest flux està compost per les dades pròpiament dites, que representen característiques d'entitats del món real, i per les metadades, dades que ofereixen informació sobre les altres dades transferides o emmagatzemades.

Les metadades estan associades a tots els components de la FIC (vegeu la figura següent), però són un component per si mateixes. Inmon les defineix dins la FIC com la "goma d'enganxar" que manté unida la resta de components, per això les considera com el component més important de la FIC.



Les metadades com a component de la FIC.

6.1.1. Metadades en les fonts de dades

Les bases de dades dels sistemes operacionals o les fonts de dades en general, des del punt de vista de la FIC, tenen com a component fonamental les dades. Però, a més d'aquestes, hi ha metadades, les quals estan generades per eines CASE (si aquestes s'han utilitzat en la seva construcció); en cas d'estar construïdes sobre un SGBD tindrem aquelles que defineixen les bases de dades que intervenen i les relacions entre els seus elements.

Generalment, en les fonts de dades les metadades descriuran, entre altres característiques, les estructures segons les quals s'emmagatzemen les dades, la quantitat de registres emmagatzemats, la seva forma d'emmagatzematge i les condicions sota les quals es produeixen les dades.

6.1.2. Metadades en els magatzems de dades

En els magatzems de dades tindrem les metadades associades als SGBD sobre els quals estan construïdes, en trobem de similars a les descrites per a les fonts de dades. A més, es pot trobar informació sobre l'ús de les dades per part dels usuaris: estadístiques d'ús, informació sobre seguretat (qui està autoritzat a fer quines operacions), etc.

6.1.3. Metadades en el component d'integració i transformació

El component d'integració i transformació utilitza les metadades de la resta de components, però, a més, pot definir com a metadades l'origen de les dades, la seva destinació, les transformacions que es fan en les dades de les fonts per a obtenir les dels magatzems i la freqüència o resultat d'aquestes transformacions.

Un cop definides totes les metadades, a partir d'aquestes es pot generar de manera automàtica el programari que faci la funció d'aquest component. És més fàcil i ràpid mantenir les metadades que mantenir un programari desenvolupat manualment[8].

[8]És a dir, si s'ha desenvolupat manualment, les metadades esmentades formarien part de la seva documentació.

Les metadades són el component més important de la FIC, ja que cohesionen la resta de components dels quals també formen part.

6.2. Ús i tipus de metadades

La informació oferta per les metadades ens permet d'entendre millor l'estructura, el funcionament i els resultats dels sistemes que descriuen. És a dir, les metadades resulten interessants per a l'equip de desenvolupament del sistema, els tècnics que fan que el sistema funcioni i els usuaris finals que l'utilitzen. Així, les podem classificar segons el paper de les persones que les fan servir. Aquests conjunts de metadades no són disjunts, és a dir, s'utilitzaran les mateixes metadades amb objectius diferents.

6.2.1. Metadades de construcció

Els equips de desenvolupament dels sistemes defineixen gran part de les metadades; posteriorment aquestes es faran servir amb altres objectius diferents en la construcció dels sistemes. En el cas de la FIC, defineixen l'estructura de les diferents fonts de dades, dels magatzems de dades, les transformacions que cal fer, la planificació, etc.

Les metadades de construcció tenen gran importància, ja que fan que els sistemes sobre els quals es defineixen siguin més flexibles i fàcils d'evolucionar.

Les metadades són tan importants en aquest aspecte dels sistemes que de vegades aquest és l'únic ús que se'ls reconeix.

6.2.2. Metadades de gestió

Durant el funcionament del sistema, per a gestionar-lo s'utilitzen algunes de les metadades definides durant la construcció i també se'n defineixen d'altres de noves. Totes aquestes formen les metadades de gestió. En la FIC es defineixen els usuaris que utilitzaran els diferents magatzems de dades, s'emmagatzema informació sobre l'ús que en fan, sobre el resultat de les extraccions i les transformacions de dades fetes, etc.

Les metadades de gestió són utilitzades pels tècnics que administren el sistema i fan que aquest funcioni.

6.2.3. Metadades d'ús

Els analistes generalment no definiran metadades (tampoc dades), almenys directament, sinó que es limiten a fer consultes sobre aquestes. A més de consultar dades, també necessiten fer consultes sobre les metadades tant de construcció com de gestió. No tindran accés a totes les metadades definides, sinó solament a aquelles que els constructors del sistema hagin considerat del seu interès segons el seu perfil d'usuari.

Exemple de metadades d'ús

Un analista necessita consultar els resultats de les vendes d'una cadena de botigues. Les vendes registrades en els sistemes operacionals de les botigues es carreguen cada dia en el magatzem de dades utilitzat. L'analista pot consultar els resultats pròpiament dits, el significat d'una dada concreta (la fórmula utilitzada per a calcular-la: metadades de construcció) i també pot fer consultes sobre incidències particulars que han tingut lloc per a obtenir-les (si falten les dades d'alguna botiga: metadades de gestió).

Generalment, els usuaris dels sistemes operacionals només necessiten treballar amb les dades de negoci emmagatzemades en els sistemes. Els usuaris dels magatzems de dades, a més de dades, necessiten metadades.

Les metadades, tant de construcció com de gestió, tenen gran importància per als usuaris dels magatzems de dades, ja que els subministren la informació que necessiten sobre el significat o l'estat de les dades que consulten.

6.3. El procés de definició de les metadades

Els avantatges de disposar de metadades en qualsevol sistema són indubtables, ja que proporcionen explícitament informació que facilita l'evolució, gestió i ús del sistema. Podem definir les metadades de manera manual o bé amb el suport d'alguna eina; així mateix, es poden definir abans, durant o després de la construcció del sistema al qual estan associats.

La situació ideal és que disposem d'una eina per a construir el sistema i que la definició de les metadades associades formi part del procés de construcció i manteniment, de manera que el sistema i les metadades associades evolucionin conjuntament.

Si no formen part necessària del procés de desenvolupament del sistema, es corre el risc que, per limitacions en el temps de desenvolupament, en el pressupost o per altres motius, les metadades no s'actualitzin amb el sistema al qual estan associades i es produeixi, així, una discordança entre les metadades i el sistema que descriuen.

Metadades associades al sistema

D'aquesta manera com es presenta a Inmon, Imhoff i Sousa (1998), s'aconsegueix que les metadades siguin completes, no siguin un element opcional, s'actualitzin automàticament cada vegada que es facin modificacions en el sistema i no requereixin un esforç addicional per a mantenir-les.

6.4. Estàndards de metadades

En sistemes complexos (com el cas de la FIC), cada component disposa de metadades. Per a definir-les s'han pogut utilitzar diferents eines de suport: eines CASE, eines de l'SGBD, eines del component d'integració i transformació, etc. Per tant, cada component té les seves metadades, emmagatzemades segons el seu criteri i format particular.

Per a poder compartir les metadades, els diferents components han "de parlar" el mateix idioma en aquest aspecte. Un estàndard de definició de metadades representa aquest idioma comú.

Lectura complementària

Podeu trobar més detalls sobre els estàndards en un annex dedicat a aquest tema a **W. A. Giovinazzo** (2000). *Object Oriented Data Warehouse Design*. Nova Jersey: Prentice Hall PTR.

Al llarg de la història (és una història relativament curta, s'inicia al principi dels noranta) s'han definit diferents estàndards relacionats amb metadades. Particularment, relacionades amb la FIC, trobem principalment el *common warehouse metadata* (CWM). El seu objectiu és definir un repositori central que permeti d'**integrar** les metadades que hi ha definides per les diferents eines, de manera que mantingui una única versió de totes aquestes. Per a això, CWM defineix un model de dades per a l'emmagatzematge format per submodels específics per a cada àrea, i un conjunt de capes d'accés al repositori que ofereixen diferents graus de funcionalitat.

6.5. Metadades històriques

Una de les característiques dels magatzems de dades és que emmagatzemen dades històriques, com hem vist en l'apartat "Característiques d'un magatzem de dades" del mòdul "Introducció a l'emmagatzematge de dades". Al llarg del temps, les estructures i altres característiques dels components de la FIC, les dades de les fonts de dades, dels magatzems de dades, les correspondències entre aquestes, les transformacions que es fan, han pogut canviar. Al costat d'aquests components també hauran canviat les metadades que les descriuen.

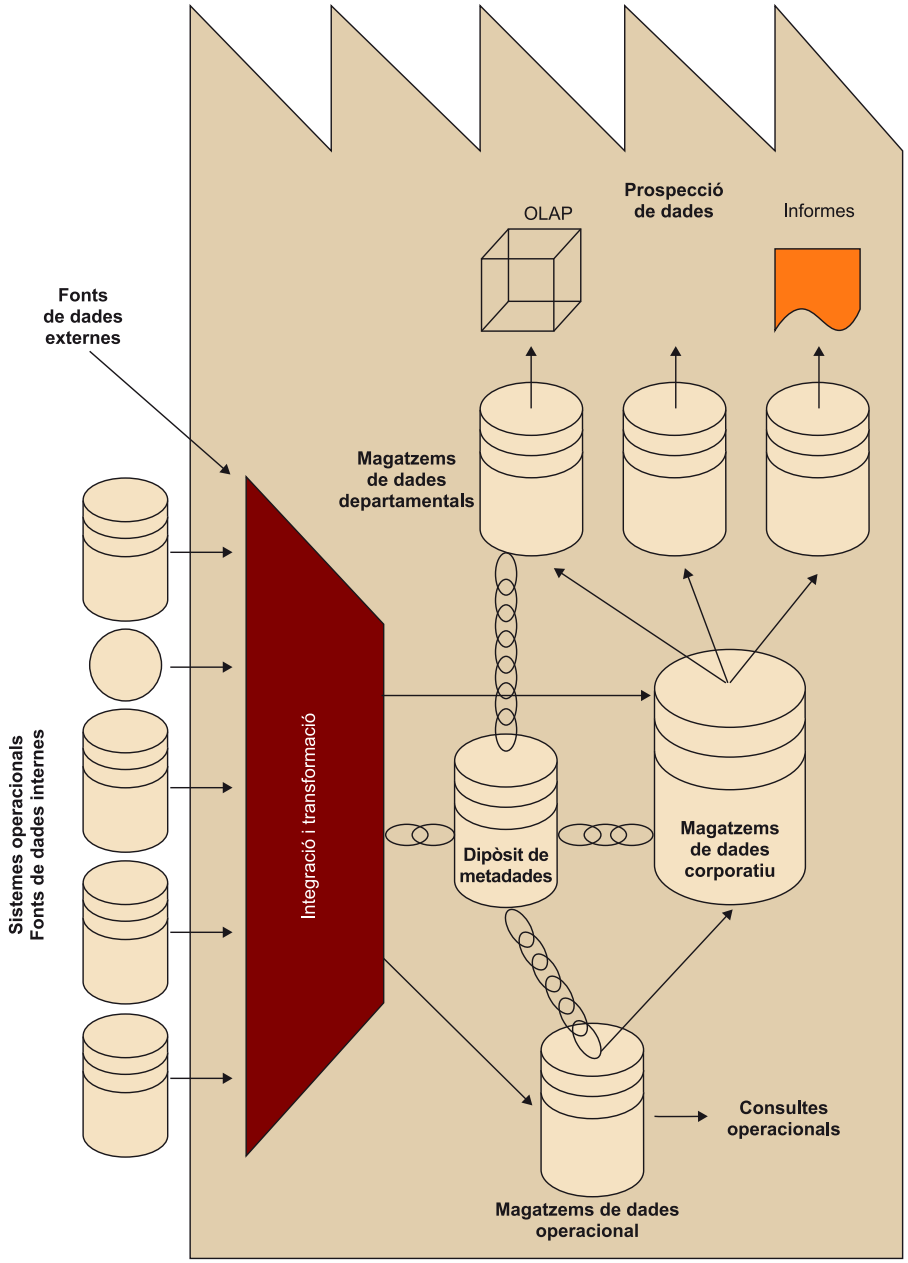
Per tant, si en els magatzems de dades tenim dades històriques amb característiques diferents, per a cada conjunt de dades definides sota les mateixes característiques haurem d'emmagatzemar la versió de les metadades que les defineixen. D'aquesta manera, els analistes podran saber per a cada dada com està emmagatzemada o en quines condicions es va obtenir.

Es necessita mantenir un control de versions de les metadades de la FIC.

7. La factoria d'informació corporativa

Quan ja hem arribat a aquest punt i coneixent de manera global els components que formen la factoria d'informació corporativa, en aquest apartat veurem com tot convergeix en un sol bloc.

La figura següent esquematitza tots els components de la factoria d'informació.



Les dades entren, provinents dels sistemes operacionals de la mateixa empresa o d'altres fonts de dades externes, directament al component d'integració i transformació. Aquest component de programari les prepara per a guardar-les

en el magatzem de dades operacional o directament en el magatzem de dades corporatiu. També és aquest component de transformació qui genera una part de les metadades que utilitzaran la resta de components en el seu funcionament. Les dades del magatzem de dades operacional serviran tant per a ser consultades, com per a alimentar el magatzem de dades corporatiu. Finalment, segons la utilitat que es donarà a les dades, aquestes es dipositen en petits magatzems de dades departamentals que estan a punt per a ser consultades o tractades.

Probablement, a causa de la joventut de l'àrea, actualment es produeix una certa confusió en els termes. Les empreses, per desconeixement o simplement per donar més importància al seu projecte, acostumen a anomenar "*data warehouse*" no al magatzem de dades corporatiu, sinó al que realment solament és un magatzem de dades departamental (és a dir, un *data mart*).

Un altre abús de terminologia també força comú és anomenar el tot (la factoria d'informació corporativa) com si fos només una part (el magatzem de dades). Es parla d'un component en comptes de parlar del procés que utilitza aquest component. Stephen R. Gardner defineix l'emmagatzematge de dades com un procés, no un producte, per a reunir i governar dades de diferents procedències amb la finalitat d'obtenir una visió única i detallada, total o parcial, d'un negoci. Aquesta idea no sembla tan diferent de la factoria d'informació presentada per William Inmon. Més aviat només és un altre punt de vista, que en certa manera inclou el primer. El fet de parlar d'un procés implica que hi hagi elements que el facin possible o, com a mínim, que ajudin a fer-lo possible.

Lectura recomanada

Podeu veure aquesta definició de l'emmagatzematge de dades a **R. S. Gardner** (1998, setembre). "Building the Data Warehouse". *Communication of the ACM* (41, 9, pàg. 52-60).

Podem considerar la factoria d'informació com el conjunt d'elements que fan possible el procés d'emmagatzematge d'informació. El magatzem de dades simplement seria un component més, com també ho són el repositori de metadades, el component d'integració i transformació, etc.

En aquest punt, encara ens podríem plantejar la necessitat d'aquesta factoria d'informació. Per què cal afegir tota aquesta complexitat als sistemes d'informació de l'empresa? Si ja tenim les dades en els sistemes operacionals, per què les repliquem en la factoria d'informació? Per què no consulten les dades directament en els sistemes operacionals els analistes? No estem malbaratant recursos? Podeu trobar les respostes a aquestes preguntes més o menys implícites en els apartats anteriors d'aquest mateix mòdul, però ara desmentirem explícitament aquesta suposada duplicitat de dades:

- Els sistemes operacionals contenen les dades que l'empresa utilitza en el seu dia a dia en l'execució del negoci. En canvi, la factoria d'informació conté dades d'anàlisi, generalment estretes d'aquests sistemes operacionals, però no necessàriament coincidents. Hi pot haver dades operacionals

(per exemple, el número de telèfon dels clients) que no interessin per a prendre decisions i dades molt importants per a prendre decisions (com el benefici) que no s'utilitzin en el funcionament diari de l'empresa.

- Generalment, els sistemes operacionals no contenen dades històriques per a no alentir innecessàriament el seu funcionament. En canvi, aquestes dades històriques són imprescindibles a l'hora de prendre decisions.

- Els sistemes operacionals sempre guarden les dades detallades (per exemple, els articles venuts a cada client). En els sistemes decisionals, de vegades, no els interessa entrar en tant de detall. El que únicament volen és l'import total de la venda, la despesa mensual del client o, simplement, el total venut durant el mes a tots els clients.

- Finalment, una altra diferència entre les bases de dades dels sistemes operacionals i les de la factoria d'informació és que aquestes últimes contenen dades netes. Durant la fase d'entrada de dades a la factoria d'informació, aquestes es netegen, se substitueixen o s'eliminen els valors nuls, es detecten inconsistències, possibles contradiccions entre diferents fonts de dades, etc. En els sistemes operacionals, amb una entrada contínua de dades, no es pot garantir aquesta netedat.

La factoria d'informació no conté les mateixes dades que els sistemes operacionals, tot i que la intersecció no és buida.

Resum

En aquest mòdul hem estudiat els diferents components que constitueixen la factoria d'informació corporativa. Hem començat pels dos extrems de la cadena (usuaris i fonts d'informació) i hem seguit amb la resta de components intermedis:

- Magatzem de dades departamental
- Magatzem de dades corporatiu
- Magatzem de dades operacional
- El component d'integració
- Les metadades

Tots aquests components s'han estudiat fent referència als sistemes operacionals, per tant, podem dir que en aquest mòdul hem completat l'estudi comparatiu entre magatzems de dades i bases de dades operacionals fet en el mòdul "Introducció a l'emmagatzematge de dades".

Finalment, hem engranat tots aquests components per a constituir l'arquitectura de la FIC. Hem tingut en compte que els components que configuren la FIC són complementaris i interactuen entre si per a satisfer les necessitats dels analistes.

Exercicis d'autoavaluació

1. Quina és la diferència principal entre el magatzem de dades corporatiu i el departamental?

2. Com justificaríeu la necessitat del magatzem de dades operacional?

3. En què es diferencien les eines OLAP de les de prospecció de dades, en relació amb les seves necessitats de dades?

4. Què vol dir que el magatzem de dades corporatiu no es pot dissenyar tenint en compte la seva funcionalitat?

5. Per què els magatzems de dades departamentals no s'alimenten directament dels sistemes operacionals, en comptes de fer-ho del magatzem de dades corporatiu?

6. Quines operacions fa el component d' integració i transformació?

7. Quin és l'element principal del component d'integració i transformació?

8. Què és el sistema de registre?

9. Quines dues fases podem distingir en l'operativa del component d'integració i transformació?

10. Quin paper tenen les metadades en la FIC?

11. Quin tipus de metadades trobem en la FIC?

12. Per què són importants els estàndards de metadades?

13. Hi ha redundància entre les dades de les bases de dades operacionals i les de la factoria d'informació corporativa?

14. Ompliu la taula següent indicant les principals diferències que hi ha entre els sistemes operacionals i decisionals:

Característica	Sistemes operacionals	Sistemes decisionals
Usuaris típics		
Nombre d'usuaris		
Tuples als quals s'ha accedit		
Objectiu del sistema		
Funcions principals		
Disseny		
Dades, característiques de		
Ús		
Accés		
Unitat de treball		
Requeriments		
Grandària		

Solucionari

1. La diferència principal és la grandària. Mentre el magatzem de dades corporatiu conté totes les dades que interessen o poden arribar a interessar a qualsevol de l'empresa, un magatzem departamental només conté aquelles que en un moment donat interessen a un cert conjunt d'analistes.

2. El magatzem de dades operacionals serveix per a satisfer eficientment i sense interferir en els sistemes operacionals les necessitats d'accés integrat a dades no històriques.

3. Les eines OLAP corresponen al que hem anomenat *grangers*, és a dir, accessos regulars a petites quantitats de dades normalment resumides per a ser mostrades als usuaris. Per contra, les eines de prospecció de dades corresponen al que hem anomenat *explorador*, accessos esporàdics a grans quantitats de dades tan detallades com sigui possible per a fer estudis estadístics.

4. El cicle de desenvolupament dels sistemes operacionals comença amb la definició dels requeriments o funcionalitat que han de donar. En canvi, el magatzem de dades corporatiu es construeix sense saber del cert quina serà la necessitat concreta que satisfarà. Per tant, es dissenya segons els temes interessants que hi hagi definits.

5. Si carreguem les dades dels magatzems de dades departamentals directament de les bases de dades operacionals, multipliquem els processos necessaris d'integració i transformació de les dades.

6. El component d'integració i transformació obté les dades de les fonts de dades, les depura, transforma i integra, les transporta als magatzems de dades i les hi carrega. També obté dades del magatzem de dades operacional i les transforma, transporta i carrega en el magatzem de dades corporatiu. A més, fa la mateixa operació entre el magatzem de dades corporatiu i els magatzems de dades departamentals.

7. A diferència d'altres components de la FIC l'element principal dels quals és la base de dades, l'element principal del component d'integració i transformació és el programari que implementa la seva missió.

8. És la font més adequada d'entre totes les fonts possibles per a les dades que s'emmagatzemen en el magatzem de dades corporatiu.

9. En una primera fase s'obtenen les dades que formen la imatge inicial de les dades. En una segona fase, iterativament s'obtenen les actualitzacions que s'han fet sobre les dades per a anar formant la "pel·lícula" (seqüència d'imatges) que ens mostra l'evolució de les dades.

10. Les metadades generalment són dades que ens donen informació sobre altres dades. En la FIC, són el component que s'encarrega de cohesionar la resta de components.

11. En trobem tres tipus, segons els usuaris que els generen o utilitzen: metadades de construcció (utilitzades per desenvolupadors), de gestió (utilitzades pels tècnics que administren i gestionen els sistemes) i d'ús (utilitzades pels analistes). Una metadada pot ser dels tres tipus, si és utilitzada per tots tres tipus d'usuari.

12. Perquè cada eina que utilitzem en la construcció de la FIC definirà les seves metadades utilitzant un format determinat. Els estàndards permetran a les diferents eines d'intercanviar i compartir metadades.

13. Sí, hi ha algunes dades que estan en tots dos sistemes. Però aquesta redundància és mínima i necessària, ja que els sistemes operacionals no guarden dades històriques, ni agregades, ni han passat un procés de neteja i integració.

14. Les diferències principals entre els sistemes operacionals i els decisionals són les següents:

Característica	Sistemes operacionals	Sistemes decisionals
Usuaris típics	Administratius	Analistes (executius)
Nombre d'usuaris	Milers	Centenars
Tuples als quals s'ha accedit	Centenars	Milers
Objectiu del sistema	Execució del negoci	Anàlisi del negoci

Característica	Sistemes operacionals	Sistemes decisionals
Funcions principals	Operacions diàries (OLTP)	Presa de decisions (OLAP)
Disseny	Orientat a la funcionalitat	Orientat al tema
Característiques-de les Dades	Actuals i actualitzades, atòmiques, normalitzades, aïllades	Històriques, resumides (agregades), desnormalitzades, integrades
Ús	Repetitiu i rutinari (consultes predeterminades)	Esporàdic i innovador (consultes *ad hoc*)
Accés	R/W	Principalment lectura
Unitat de treball	Transaccions simples	Consultes complexes
Requeriments	Rendiment de transaccions + consistència de dades	Rendiment de les consultes i precisió de les dades
Grandària	MB/GB	GB/TB

Glossari

dada (definició des del punt de vista dels sistemes decisionals) *f* Mesura, observació feta i emmagatzemada en algun sistema.

factoria d'informació corporativa *f* Conjunt d'elements de programari i maquinari que ajuden en l'anàlisi de dades per a prendre decisions.
sigla: FIC

FIC *f* Vegeu **factoria d'informació corporativa**.

informació (definició des del punt de vista dels sistemes decisionals) *f* Dades rellevants per a algú que decideix, que afecten alguna de les seves decisions.

magatzem de dades corporatiu *m* Conjunt de dades que guarda integrades totes les dades històriques de l'empresa.

magatzem de dades departamental *m* Conjunt de dades que resol les necessitats d'anàlisi d'un cert departament o conjunt d'usuaris.

magatzem de dades operacional *m* Conjunt de dades integrat i orientat al tema, però sense dades històriques. S'acostuma a utilitzar com a pas intermedi en la construcció del magatzem de dades corporatiu.

metadada *f* Dades sobre dades.

OLAP Sigles que fan referència a les eines d'anàlisi, normalment multidimensional.
en on-line analytical processing

OLTP *On-line transaccional processing*.

SGBD *m* Vegeu **sistema de gestió de bases de dades**.

sistema de gestió de bases de dades *m* Programari que gestiona i controla bases de dades. Les seves funcions principals són les de facilitar-ne l'ús simultani a molts usuaris de tipus diferents, independitzar l'usuari del món físic i mantenir la integritat de les dades.
sigla: SGBD
en database management system

sistema de registre *m* Font de cadascuna de les dades dels magatzems de dades, d'entre totes les fonts possibles.

sistema operacional *m* Aquell que ajuda en les operacions diàries del negoci d'una organització.

sistema transaccional *m* Aquell basat en transaccions de lectura/escriptura.

Bibliografia

Asociación de Técnicos en Informática (1999, març-abril). *Novática* (núm. 138).

Giovinazzo, W. A. (2000). *Object Oriented Data Warehouse Design*. Nova Jersey: Prentice Hall PTR.

Inmon, W. H.; Imhoff, C.; Sousa, R. (1998). *Corporate Information Factory*. EUA: John Wiley & Sons, Inc.

Jarque, M.; Lenzerini, M.; Vassiliou, Y.; Vassiliadis, P. (2000). *Fundamentals of Data Warehouses*. Berlín: Springer-Verlag.